CISC839 G-13 Final Report: Mining Airbnb Rental Data

Ahmed Mohamed Gaber[1], Karim Gamal Mahmoud[2], and Sara Ahmed Mohamed[3]

[1]**Email : 21amga@queensu.ca**
[2]**Email : 21kgmm@queensu.ca**
[3]**Email : 21sama2@queensu.ca**

1 BACKGROUND AND OBJECTIVE

People are more flexible than ever about where and when they travel. To help them take advantage of these new possibilities, Airbnb are introducing biggest change in a decade—including a completely new way to search, a better way to stay longer, and an unmatched level of protection. the main objective of this project is to predict the cost of listing to help user who want to post their houses to specify the appropriate price.

Through the analysis of this project, we are going to answer some of the potential question that will help us in achieving our objective, such as:

- **hypothesis test question**

  ¿ The average of listings price with a rating of 4.5 or higher are more expensive than those with a lower rating!

  ¿ This will be useful for people who are looking for cheap prices for real estate with a good rating.

- **regression**

  ¿ Can we predict the price of the listings based on available features?

  ¿ It will benefit for customers to find the best listings price based on their features

- **predictive analysis**

  ¿ Can we predict the price level of the listings after converting the prices to three levels(categories): low, medium, high.

  according to: ’prices that are less or equal to 25% as (low), 25% to 75% as (medium), and 75% or higher as (high)’ ?

  ¿ This method will make it easier for user to find the suitable listings based on the category of the listings price (the price range)

2 DATASET

Airbnb provides open data that quantifies the impact of short-term rentals on housing and residential communities; and also provides a platform to support advocacy for policies to protect our cities from the impacts of short-term rentals. By analyzing publicly available information about a city’s Airbnb’s listings, Inside Airbnb provides filters and key metrics so we can see how Airbnb is being used to compete with the residential housing market.

The connection between the multiple files existed in the provided dataset is ’listing$_i d'column.$

Description of Data:

The dataset comprises of three main tables:

- listings - Detailed listings data showing 96 attributes for each of the listings. Some of the attributes used in the analysis are price(continuous), longitude (continuous), latitude (continuous), listing_type (categorical), is_superhost (categorical), neighbourhood (categorical), ratings (continuous) among others.
- reviews - Detailed reviews given by the guests with 6 attributes. Key attributes include date (datetime), listing_id (discrete), reviewer_id (discrete) and comment (textual).
- calendar - Provides details about booking for the next year by listing. Four attributes in total including listing_id (discrete), date(datetime), available (categorical) and price (continuous).

2.1 Data Preprocessing

2.1.1 Removing Duplicates

- Check if there is any duplicate rows and remove them.

2.1.2 Handling Missing Values

- Remove columns that has nulls more than 30%.
- Remove records.
- Using imputation (IterativeImputer, SimpleImputer).

2.1.3 Handling Outliers

- Replacing by the mean or median.
- Removing outliers.

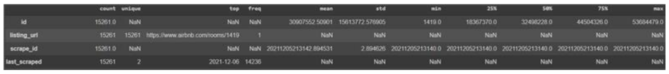
2.1.4 Handling Datatyps

- Convert categorical features to numerical.

2.2 Basic Statistics of the Dataset

2.2.1 Listings

Data shape: before preprocessing = 15261 rows × 74 columns, and after preprocessing = 12060 rows × 1678 columns
Data Duplication: the data doesn't contain any duplicates
Missing Data: the dataset contain many missing data in more than 40 columns

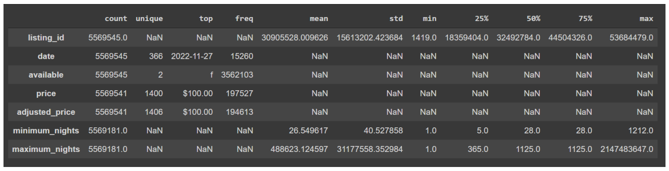


	count	unique	top	freq	mean	std	min	25%	50%	75%	max
id	15261.0	NaN	NaN	NaN	30907552.50901	15613772.576905	1419.0	18367370.0	32498228.0	44504326.0	53684479.0
listing_url	15261	15261	https://www.airbnb.com/rooms/1419	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
scrape_id	15261.0	NaN	NaN	NaN	20211205213142.894531	2.894626	20211205213140.0	20211205213140.0	20211205213140.0	20211205213140.0	20211205213140.0
last_scraped	15261	2	2021-12-06	14236	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 1. Listings description

2.2.2 Calendar

Data shape: before preprocessing = 5569545 rows × 7 columns, and after preprocessing = 4746530 rows × 7 columns
Data Duplication: the data doesn't contain any duplicates
Missing Data: the dataset contain very small amount of missing data



	count	unique	top	freq	mean	std	min	25%	50%	75%	max
listing_id	5569545.0	NaN	NaN	NaN	30905528.009626	15613202.423684	1419.0	18359404.0	32492784.0	44504326.0	53684479.0
date	5569545	366	2022-11-27	15260	NaN	NaN	NaN	NaN	NaN	NaN	NaN
available	5569545	2	f	3562103	NaN	NaN	NaN	NaN	NaN	NaN	NaN
price	5569541	1400	$100.00	197527	NaN	NaN	NaN	NaN	NaN	NaN	NaN
adjusted_price	5569541	1406	$100.00	194613	NaN	NaN	NaN	NaN	NaN	NaN	NaN
minimum_nights	5569181.0	NaN	NaN	NaN	26.549617	40.527858	1.0	5.0	28.0	28.0	1212.0
maximum_nights	5569181.0	NaN	NaN	NaN	488623.124597	31177558.352984	1.0	365.0	1125.0	1125.0	2147483647.0

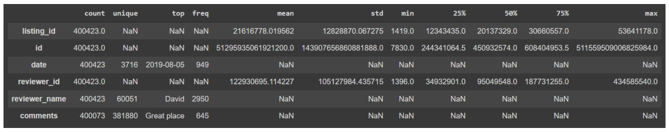
Figure 2. Calendar description

2.2.3 *Reviews*

Data shape: before preprocessing = 400423 rows × 6 columns, and after preprocessing = 400073 rows × 3 columns

Data Duplication: the data doesn’t contain any duplicates

Missing Data: the dataset contain very small amount of missing data 0.087%



	count	unique	top	freq	mean	std	min	25%	50%	75%	max
listing_id	400423.0	NaN	NaN	NaN	21616778.019562	12828870.067275	1419.0	12343435.0	20137329.0	30660557.0	53641178.0
id	400423.0	NaN	NaN	NaN	51295935061921200.0	143907656860881888.0	7830.0	244341064.5	450932574.0	608404953.5	511559509006825984.0
date	400423	3716	2019-08-05	949	NaN	NaN	NaN	NaN	NaN	NaN	NaN
reviewer_id	400423.0	NaN	NaN	NaN	122930695.114227	105127984.435715	1396.0	34932901.0	95049548.0	187731255.0	434585540.0
reviewer_name	400423	60051	David	2950	NaN	NaN	NaN	NaN	NaN	NaN	NaN
comments	400073	381880	Great place	645	NaN	NaN	NaN	NaN	NaN	NaN	NaN

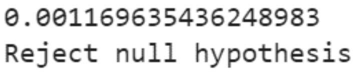
Figure 3. Reviews description

3 ANSWERS TO THE RESEARCH QUESTIONS

3.1 Answer to question #1: The average of listings price with a rating of 4.5 or higher are more expensive than those with a lower rating!

According to the Approach, The statistical test we used is T-test (using stats.ttest_ind), We can use this test, if we observe two independent samples from the same or different population. The test measures whether the average (expected) value differs significantly across samples. If we observe a large p-value, for example larger than 0.05, then we cannot reject the null hypothesis of identical average scores. If the p-value is smaller than the threshold, then we reject the null hypothesis of equal averages.

The result is:



```
0.001169635436248983
Reject null hypothesis
```

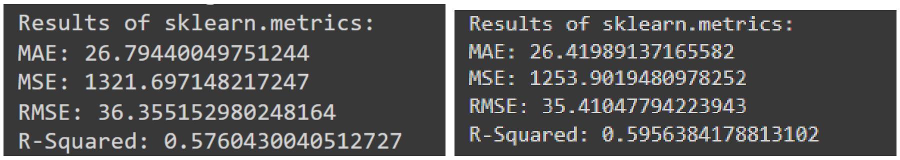
Figure 4. Hypothesis Test Result

From the result we can notice that The hypothesis test failed as the p-value is small than the threshold which mean that rating of listing affect its price and the average of listings price with a rating of 4.5 or higher are more expensive than those with a lower rating.

3.2 Answer to question #2: Can we predict the price of the listings based on available features?

- The approach that we used:
  1- We used RandomForestRegressor to select the best features (features selection) based on the price
  2- Selecting the top 20 features correlated with price
  3- We run the selected features on these models: [XGBRegressor, DecisionTreeRegressor, RandomForestRegressor]
- For hyperparameters, we used the default hyperparameter for all the models
- We split the data into 80% train and 20% test
  For encoding our input, we used get_dummies to convert the input text from categorical to numerical

- The best two models results:



```
Results of sklearn.metrics:
MAE: 26.79440049751244
MSE: 1321.697148217247
RMSE: 36.355152980248164
R-Squared: 0.5760430040512727
```

```
Results of sklearn.metrics:
MAE: 26.41989137165582
MSE: 1253.9019480978252
RMSE: 35.41047794223943
R-Squared: 0.5956384178813102
```

Figure 5. RandomForest VS XGBRegressor Result

- As we can see from the results based on MAE and R-Squared, we found that XGBRegressor is better than RandomForest because the r-squared for XGBRegressor is higher than RandomForest, at the same time the MAE for XGBRegressor is less than RandomForest

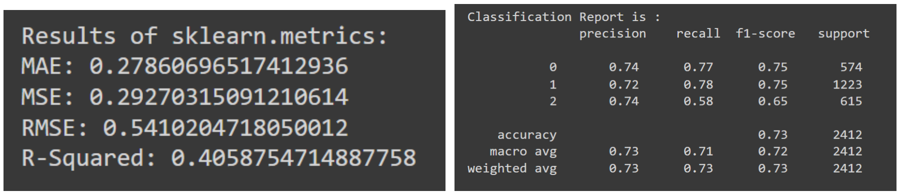
3.3 Answer to question #3: Can we predict the price level of the listings after converting the prices to three levels(categories): low, medium, high. according to: 'prices that are less or equal to 25% as (low), 25% to 75% as (medium), and 75% or higher as (high)' ?

- The approach that we used:
  1- Create price category column that contain only three values: Low, Medium, High
  2- Convert categorical column to numerical (Low = 0, Medium = 1, High = 2)
  3- We used RandomForestClassifier to select the best features (features selection), then we run the selected features on these models: [LogisticRegression, XGBClassifier, DecisionTreeClassifier, RandomForestClassifier]

- For hyperparameters, we used the default hyperparameter for all the models

- We split the data into 80% train and 20% test
  For encoding our input, we used get_dummies to convert the input text from categorical to numerical

- The best two models results:



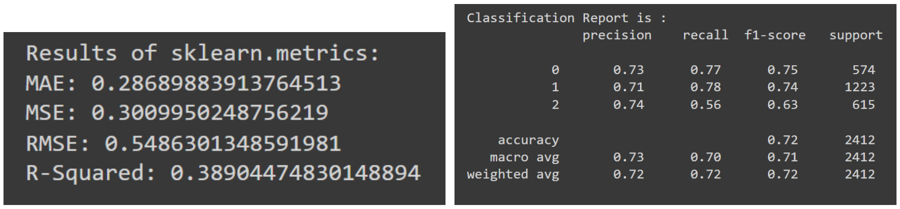
```
Results of sklearn.metrics:
MAE: 0.27860696517412936
MSE: 0.29270315091210614
RMSE: 0.5410204718050012
R-Squared: 0.4058754714887758
```

```
Classification Report is :
              precision    recall  f1-score   support

           0       0.74      0.77      0.75       574
           1       0.72      0.78      0.75      1223
           2       0.74      0.58      0.65       615

    accuracy                           0.73      2412
   macro avg       0.73      0.71      0.72      2412
weighted avg       0.73      0.73      0.73      2412
```

Figure 6. RandomForest Result



```
Results of sklearn.metrics:
MAE: 0.28689883913764513
MSE: 0.3009950248756219
RMSE: 0.5486301348591981
R-Squared: 0.38904474830148894
```

```
Classification Report is :
              precision    recall  f1-score   support

           0       0.73      0.77      0.75       574
           1       0.71      0.78      0.74      1223
           2       0.74      0.56      0.63       615

    accuracy                           0.72      2412
   macro avg       0.73      0.70      0.71      2412
weighted avg       0.72      0.72      0.72      2412
```

Figure 7. XGBClassifier Result

- As we can see from the results based on MSE and the accuracy, we found that RandomForest is better than XGBClassifier and this maybe occured because we used the default hyperparamters
- And we discovered that the reason for this low accuracy number for XGB model because we used different hyperparameters on the other hand we used the default hyperparameters for Random Forest model which gives us a higher accuracy
- **We didn't use the AUC curve because we were dealing with multiple classes and the AUC doesn't work with that**

4 LIMITATIONS

4.1 Huge Data

We have encountered limitations in the large size of the calendar data, although the size of the data was smaller after preprocessing, but we had almost 5 million records, which make us to train the decision tree model on 50% of the data not 85% or 90% as usual.

We also faced a problem with the large number of features for Listing dataset, which were 74 features, and there was not enough information to facilitate the preprocessing of them even on the Airbnb official website .

4.2 Generalization

This analysis wouldn't be generalized to locations other than Toronto. Also, as the time going by, this analysis may not be very effective

5 TAKE-AWAY MESSAGES

5.1 Conclusion

- What we discovered is that the price prediction will be more accurate using 30 of the listings features.
- What we discovered is that the price category classification will be more accurate using 40 of the listings features
- We noticed that the number of hosts whose prices are medium is twice the number of hosts whose prices are high and cheap.
- We conclude from this that most of the prices in Toronto are medium prices
- rating of listing affect its price and the average of listings price with a rating of 4.5 or higher are more expensive than those with a lower rating!

5.2 future potential implications

- It will help people to choose better hosts with better prices by extracting the best features by using our the prediction models.
- It will also help people to search for the best listings that fit their budget according to the price category

6 REPLICATION PACKAGE

This is a Google colaboratory notebook to run the entire project using the GPU and data uploaded to our google drive.

- Colab: Colab Link
- GitHub: Github Link

7 DISTRIBUTION OF WORKLOAD

Data Preprocessing:-

- Ahmed and Karim : worked on (listings, calendar, neighbourhoods files)
- Sara : worked on (reviews file)

Data Analysis:-

- Ahmed and Karim : worked on (listings, calendar, neighbourhoods files)
- Sara : worked on (reviews file)

Feature Engineering:-

- All of us worked on it

Answering Analysis Qustions:-

- Ahmed : worked on (predictive analysis)
- Karim : worked on (regression analysis)
- Sara : worked on (hypothesis test)

Models:-

- Ahmed and Karim : worked on (RandomForestRegressor, LogisticRegression, XGBClassifier, XGBRegressor, RandomForestClassifier)
- Sara : worked on (DecisionTreeRegressor, DecisionTreeClassifier)

8 OVERLEAF LINK

OverLeaf URL : OverLeaf Link